Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction

Petra Friederichs*

Meteorological Institute, University of Bonn, Germany

Thordis L. Thorarinsdottir

Norwegian Computing Center, Oslo, Norway

Abstract

Predictions of the uncertainty associated with extreme events are a vital component of any prediction system for such events. Consequently, the prediction system ought to be probabilistic in nature, with the predictions taking the form of probability distributions. This paper concerns probabilistic prediction systems where the data is assumed to follow either a generalized extreme value distribution (GEV) or a generalized Pareto distribution (GPD). In this setting, the properties of proper scoring rules which facilitate the assessment of the prediction uncertainty are investigated and closed-from expressions for the continuous ranked probability score (CRPS) are provided. In an application to peak wind prediction, the predictive performance of a GEV model under maximum likelihood estimation, optimum score estimation with the CRPS, and a Bayesian framework are compared. The Bayesian inference yields the highest overall prediction skill and is shown to be a valuable tool for covariate selection, while the predictions obtained under optimum CRPS estimation are the sharpest and give the best performance for high thresholds and quantiles.

Keywords: Bayesian variable selection, continuous ranked probability score, extreme events, optimum score estimation, prediction uncertainty, wind gusts

1 Introduction

Extreme events in weather and climate such as high wind speeds, heavy precipitation or extremal temperatures are commonly associated with high impacts on both environment and society. However, the physical processes leading to the extremes are usually generated on small scales and their

*Corresponding author address:
Petra Friederichs, Meteorological Institute, University of Bonn, Auf dem Huegel 20, 53121 Bonn, Germany
E-mail: pfried@uni-bonn.de

prediction is contaminated by large uncertainty. The need to determine uncertainties in the predictions of extreme events is stressed in the report from a recent workshop on extreme events in climate and weather (Guttorp and Fuentes, 2010). A prediction system for such events should therefore be probabilistic in nature, allowing for an assessment of the associated uncertainty (Dawid, 1984; Gneiting, 2008).

The verification methods applied to these systems should thus necessarily be equipped to also handle the verification of uncertainty estimates. Murphy (1993) argues that a general prediction system should strive to perform well on three types of goodness: there should be consistency between the forecaster's judgment and the forecast, there should be correspondence between the forecast and the observation, and the forecast should be informative for the user. On a similar note, Gneiting et al. (2007) state that the goal of probabilistic forecasting should be to maximize the sharpness of the predictive distribution subject to calibration. Here, calibration refers to the statistical consistency between the predictive distribution and the observation, while sharpness refers to the concentration of the predictive distribution; the sharper the forecast, the higher information value will it provide. The prediction goal of Gneiting et al. (2007) is thus equivalent to Murphy's second and third type of goodness.

Verification methods that aim to attain these goals have been extensively studied in the literature, see e.g. Wilks (2011, Chapter 8) for an excellent overview. In this paper, we focus on the prediction of extreme events and our main objective is to assess the characteristics of proper scoring rules for extreme value distributions. The framework of proper scoring rules can also be used for the parameter estimation in that a scoring rule is optimized over the training data, see e.g. Gneiting et al. (2005). Optimum score estimation returns unbiased parameter estimates (Dawid, 2007) and for the ignorance score, this equals maximum likelihood estimation for independent observations. The continuous ranked probability score or CRPS (Unger, 1985; Hersbach, 2000; Gneiting and Raftery, 2007) is of particular interest in our context, as it simultaneously assesses all of Murphy's types of goodness. A closed form expression of the CRPS has been calculated for a normal distribution (Gneiting et al., 2005), for a mixture of normals (Grimit et al., 2006), for a truncated normal distribution (Gneiting et al., 2006), and for the three parameter two-piece normal distribution (Thorarinsdottir and Gneiting, 2010). We derive closed-form expressions for the CRPS for the generalized extreme value distribution (GEV) and the generalized Pareto distribution (GDP).

In an application to peak wind prediction, we compare minimum CRPS estimation, maximum likelihood estimation, and a Bayesian approach under the GEV using predictive performance as the comparison criteria. Peak winds of short duration (few seconds) are a major cause of wind-related damage and crucial in wind hazard studies. Observations of peak winds are, however, sparse since only a small proportion of the weather stations provide peak wind speed observations. Using data from the observation station Valkenburg in the Netherlands, we investigate how observations of other weather variables, including mean wind speed, precipitation, and pressure, may be used as covariates to obtain a predictive distribution for the peak wind speed. Furthermore, we show how a Bayesian regression variable selection method can simplify the covariate selection procedure when the space of potential models is large.

The remainder of the paper is organized as follows. In section 2, we discuss how the predictive performance of probabilistic forecasts for extreme events may be assessed and provide the closed form expressions for the CRPS under the GEV and the GPD. A detailed description of the derivation is given in the appendix. Our case study on probabilistic forecasting and forecast verification

for peak wind speed is presented in section 3. The paper closes with a discussion in section 4.

2 Assessing predictive performance

Murphy and Winkler (1987) and Murphy (1993) define a general framework for deterministic forecast verification based on two aspects of prediction quality: reliability and resolution. Reliability, or calibration, relates to the correspondence between the observation and the forecast. In our probabilistic setting, a forecast F is perfectly calibrated if the conditional distribution of the observation Y given the forecast is equal to F. Resolution is a measure of information content which is closely related to the sharpness of the forecast. In our setting, sharpness is defined with respect to the predictive distribution, where a sharp predictive distribution reflects high information content or entropy. Note that sharpness has to be conditioned on F being calibrated for it to be related to the resolution – otherwise it is merely an attribute of the forecast distribution. In the following, we discuss various methods to assess the calibration, or the reliability, and the resolution of a predictive distribution for data assumed to follow an extreme value distribution as in (7) or (8) below.

2.1 Calibration

Let $F_1, \ldots, F_n$ denote the respective predictive distributions for the observations $y_1, \ldots, y_n$. A standard method to assess the calibration of the forecasts is to apply the probability integral transform (PIT) (Dawid, 1984). The PIT is given by the value of the predictive distribution at the observation, $F_i(y_i)$. If the forecasts are calibrated, the sample $\{F_i(y_i)\}_{i=1}^n$ should follow the standard uniform distribution. This may e.g. be assessed graphically through a histogram, where a $\cup$-shaped histogram will indicate that the forecasts are underdispersive, while a $\cap$-shaped histogram will indicate that the forecasts are overdispersive, see e.g. Wilks (2011). The PIT histogram is the continuous counterpart of the verification rank histogram or the Talagrand diagram (Anderson, 1996; Hamill and Colucci, 1997).

The PIT values might also be displayed in terms of a PP-plot, where they are compared to n equally spaced percentiles of the standard uniform distribution. In extreme value theory, a very popular assessment of the calibration of a non-stationary extreme value model is the so-called residual quantile plot (Coles, 2001). Residual quantile plots emphasize potential deviations in the upper tail of the distribution. To this end, the PIT values and the uniform percentiles are transformed to quantiles through some inverse distribution function G^{-1}. In general, the inverse Gumbel is used for G^{-1}, as it is the standard reference distribution for models of block maxima. For threshold models an exponential distribution is often used.

The calibration might also be estimated over a restricted range of forecasts, for instance situations with high forecast probability of extreme events. A stratified PIT histogram or residual quantile plot would then be evaluated based on $\{F_i(y_i)\}_{i \in I}$ for some $I \subset \{1, \ldots, n\}$. An important aspect here is that the subset I has to be chosen independently of the observations $y_1, \ldots, y_n$, as the observed value is not known to the forecaster when a forecast is issued.

2.2 Scoring rules

A variety of scores exist to assess the quality of a probabilistic forecast. An important characteristic of a qualified scoring rule is its propriety (Murphy, 1973; Gneiting and Raftery, 2007; Bröcker and Smith, 2007). A scoring rule is (strictly) proper if the expected score for an observation Y is optimized if (and only if) the true distribution of Y is issued as the forecast. Propriety will encourage honesty and prevent hedging which coincides with Murphy's first type of goodness (Murphy, 1993). We will only consider proper scoring rules in the following. Furthermore, the scoring rules are negatively oriented such that a lower value means a better score.

A widely used scoring rule is the logarithmic score proposed by Good (1952) which in meteorological applications is known under the name ignorance score (Roulston and Smith, 2002). It is defined as

$$S_{IGN}(f, y) = -\log(f(y)dy).$$

The ignorance score applies to the predictive density function f and is proportional to the log-likelihood of the data with respect to the predictive density. Note, that dy is ignored when calculating the log-likelihood. However, it must be taken into account if transformed variables are compared. The associated expected score is the Shannon entropy, and the divergence function becomes the Kullback-Leibler divergence (Gneiting and Raftery, 2007). It thus represents a score which is motivated by information theory. Both Selten (1998) and Grimit et al. (2006) have pointed out that, although the ignorance score is simple to calculate, it attributes a very strong penalty to events with low probability and is thus very sensitive to outliers.

The continuous ranked probability score (CRPS) assesses the predictive skill of a forecast in terms of the entire predictive distribution F (Unger, 1985; Hersbach, 2000) and can thus assess both the calibration and the sharpness of the forecast simultaneously. It is defined as

$$S_{CRP}(F, y) = \int_{-\infty}^{\infty} \left[F(t) - H(t - y)\right]^2 dt \tag{1}$$

and compares the forecast distribution F and the empirical distribution of the observation y. Here, $H(t-y)$ denotes the Heaviside step function using the half-maximum convention with $H(0) = 0.5$. Note that an observation error might easily be introduced in the CRPS, e.g. assuming Gaussian errors and using a Gaussian distribution instead of the Heaviside step function.

The CRPS can be decomposed into a reliability and a resolution part (Hersbach, 2000; Gneiting and Raftery, 2007). That is, we can write

$$S_{CRP}(F, y) = \mathbb{E}|X - y| - \frac{1}{2}\mathbb{E}|X - X'|, \tag{2}$$

where X and X' are independent random variables with distribution F. In principle, any (strictly) proper scoring rules may be decomposed into an uncertainty, a reliability and, a resolution part (Gneiting and Raftery, 2007; Bröcker, 2009). This representation of the CRPS is especially useful if a closed form expression of the CRPS is not available for F. Let $\mathbf{x}$ and $\mathbf{x}'$ denote two independent samples of size m from the predictive distribution F. The representation in (2) can then easily be approximated by

$$S_{CRP}(F, y) \approx \sum_{i=1}^{m} |x_i - y| - \frac{1}{2} \sum_{i=1}^{m} |x_i - x_i'|. \tag{3}$$

Even though the forecast is given by a full predictive distribution F, it is often of interest to focus on the prediction of certain events, such as the probability of threshold exceedance. We can assess the forecasters ability to predict events over a given threshold u with the Brier score,

$$S_B^u(F, y) = (p_u - \mathbb{1}\{y \geq u\})^2, \tag{4}$$

where $p_u = 1 - F(u)$ is the predicted probability of the realized value being greater or equal to the threshold u (Brier, 1950). Note that the CRPS in (1) represents an integral of the Brier score over all possible thresholds.

Similarly, we might want to focus on the predictive performance in the upper tail. This can be achieved by using the quantile score

$$S_Q^\tau(F, y) = \rho_\tau(y - F^{-1}(\tau)), \tag{5}$$

where y is the event that materializes, τ is the quantile of interest, $\rho_\tau(u) = \tau u$ if $u \geq 0$, and $\rho_\tau(u) = (\tau - 1)u$ otherwise (Gneiting and Raftery, 2007; Friederichs and Hense, 2007). A third alternative deviation of the CRPS is obtained using the quantile score,

$$S_{CRP}(F, y) = 2 \int_0^1 \rho_\tau(y - F^{-1}(\tau)) d\tau, \tag{6}$$

see Laio and Tamea (2007), Gneiting and Raftery (2007), and Gneiting and Ranjan (2011).

When comparing various forecasters, it might often be advantageous to compare skill scores rather than the scores themselves (Murphy, 1973, 1974). A skill score measures the relative gain of a forecast with respect to the reference forecast and it is defined as

$$SS(F, y) = \frac{S(F, y) - S(F_{ref}, y)}{S(F_{perfect}, y) - S(F_{ref}, y)}$$

where F^{ref} is a reference forecast used for all forecasters, and $F_{perfect}$ is the perfect forecast. In the case of CRPS and ignorance score the respective score of a perfect forecast is zero. A zero skill score represents no gain in predictive skill, while a perfect forecast would have a skill score of 100%. This approach is especially useful when comparing very high quantiles under the quantile score in (5) or very high thresholds under the Brier score in (4) as these will be based on relatively small amount of data.

2.3 Scoring rules for extreme value distributions

Events are generally classified as being extreme by one of two criteria: the extremes are either given by a block maxima or they are defined as all values that exceed a given threshold. The asymptotic behavior of the first type can be modeled by the generalized extreme value distribution (GEV),

$$F_{GEV}(y) = \begin{cases} \exp\big(-(1+\xi\frac{y-\mu}{\sigma})^{-1/\xi}\big), & \xi \neq 0 \\ \exp\big(-\exp(-\frac{y-\mu}{\sigma})\big), & \xi = 0 \end{cases}, \tag{7}$$

where $1+\xi\frac{y-\mu}{\sigma} > 0$ for $\xi \neq 0$ (Coles, 2001; Beirlant et al., 2004). The three parameters of the GEV are location μ, scale σ and shape ξ. For threshold excesses, the generalized Pareto distribution (GPD) is usually applied. It is given by

$$F_{GPD}(y) = \begin{cases} 1-(1+\xi\frac{y-u}{\sigma_u})^{-1/\xi}, & \xi \neq 0 \\ 1-\exp(-\frac{y-u}{\sigma_u}), & \xi = 0 \end{cases}, \tag{8}$$

where u is a sufficiently large threshold, σ_u is the scale and ξ the shape parameter.

For $\xi \neq 0$, a closed form expression of the CRPS for the GEV is given by

$$\begin{aligned} S_{CRP}(F_{GEV_{\xi\neq0}}, y) =& \Big[\mu - y - \frac{\sigma}{\xi}\Big]\Big[1 - 2F_{GEV_{\xi\neq0}}(y)\Big] \\ &- \frac{\sigma}{\xi}\Big[2^{\xi}\Gamma(1-\xi) - 2\Gamma_l\big(1-\xi, -\log F_{GEV_{\xi\neq0}}(y)\big)\Big], \end{aligned} \tag{9}$$

where Γ denotes the gamma function and Γ_l the lower incomplete gamma function. For $\xi = 0$, this expression reads

$$S_{CRP}(F_{GEV_{\xi=0}}, y) = \mu - y + \sigma[C - \log 2] - 2\sigma Ei\big(\log F_{GEV_{\xi=0}}(y)\big), \tag{10}$$

where $C \approx 0.5772$ is the Euler-Mascheroni constant and $Ei(x) = \int_{-\infty}^{x} \frac{e^t}{t} dt$ is the exponential integral (Abramowitz and Stegun, 1964). The derivations are given in Appendix A. In the case of the GPD, we get

$$\begin{aligned} S_{CRP}(F_{GPD_{\xi\neq0}}, y) =& \Big[u - y - \frac{\sigma_u}{\xi}\Big]\Big[1 - 2F_{GPD_{\xi\neq0}}(y)\Big] \\ &- \frac{2\sigma_u}{\xi(\xi-1)}\Big[\frac{1}{\xi-2} + \Big[1 - F_{GPD_{\xi\neq0}}(y)\Big]\Big[1 + \xi\frac{y-u}{\sigma_u}\Big]\Big] \end{aligned} \tag{11}$$

for $\xi \neq 0$ and

$$S_{CRP}(F_{GPD_{\xi=0}}, y) = y - u - \sigma_u\Big[2F_{GPD_{\xi=0}}(y) - \frac{1}{2}\Big] \tag{12}$$

for $\xi = 0$. The derivations for a GPD are given in Appendix B.

Figure 1 shows the CRPS, the ignorance score and the quantile score for $\tau = 0.5$ for the GEV with shape parameter $\xi = -0.5$ (Weibull type), $\xi = 0$ (Gumbel type), and $\xi = 0.5$ (Fréchet type), as well as for the GDP with the same parameter values. The ignorance score has a sharper minimum than the other scores and takes its minimum at the mode of the predictive distribution while the other two scores take their minimum at the median of the distribution. For the CRPS under the GEV, this can easily be shown by introducing the median, $F_{GEV}(0.5) = \mu + \frac{\sigma}{\xi}((\log 2)^{-\xi} - 1)$ for $\xi \neq 0$ and $F_{GEV}(0.5) = \mu - \sigma\log(\log 2)$ for $\xi = 0$, into the derivative of the score with respect to y.

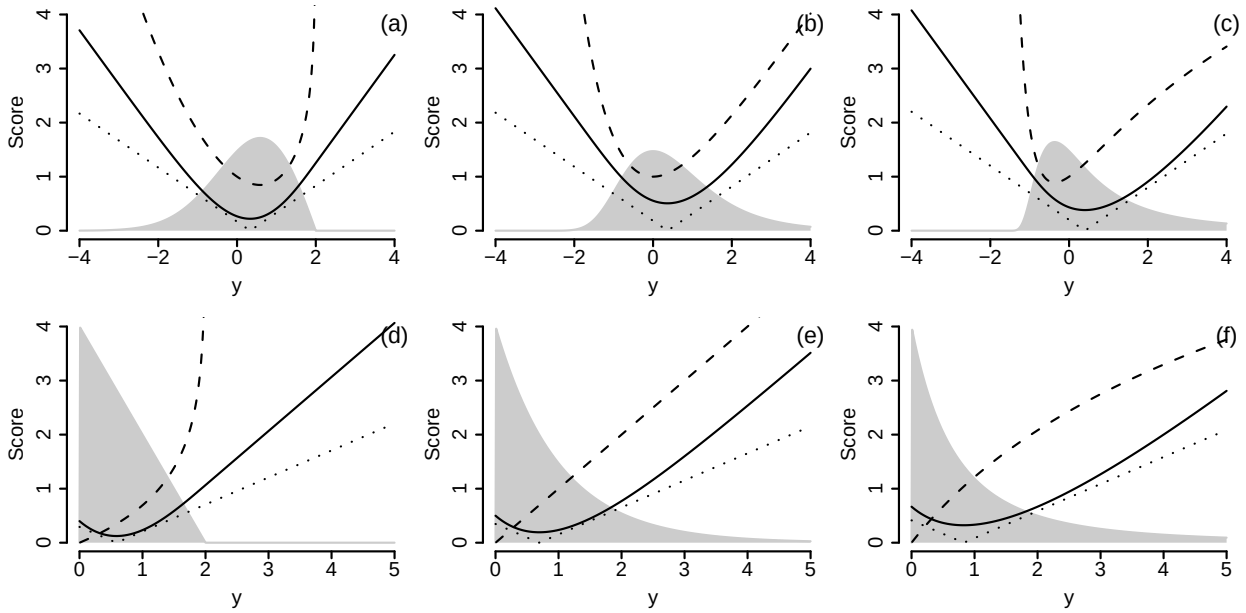

Figure 1: The CRPS (solid lines), the ignorance score (dashed lines), and the quantile score for $\tau = 0.5$ (dotted lines) as a function of the observation value for the GEV (top row) and the GPD (bottom row). The predictive distributions have zero location ($\mu = 0$), standard scale ($\sigma = 1$), and shape $\xi = -0.5$ (first column), $\xi = 0$ (second column), or $\xi = 0.5$ (third column). The corresponding predictive densities are indicated in gray.

3 Predicting peak wind speed

We apply the verification measures discussed in the previous sections to predictions of daily peak wind speed at a location in the Netherlands over a period from January 1, 2001 to July 1, 2009. The goal is to derive a prediction model for daily peak wind speed using the observations of other meteorological variables as covariates with out-of-sample data used for the parameter estimation. In our application, the covariates are observed at the same time and location as the observation. We thus derive nowcasts rather than real forecasts in time for daily peak wind speed. In view of the sparse observational network for gust observations, such an analysis can be very useful, e.g. for forecast verification of wind gust warnings (Friederichs et al., 2009). Furthermore, the covariates might easily be replaced by the corresponding outputs from numerical weather prediction (NWP) models, thereby providing a probabilistic methodology for model output statistics, as NWP models only give diagnostic estimates for peak wind speed.

For the modeling, we use generalized extreme value distributions under both a Bayesian forecasting framework and a frequentist framework where the parameters are estimated by minimizing a proper scoring rule. That is, we assume that peak wind speed observations follow a non-stationary GEV (7) where the parameters depend on covariates $\mathbf{x}$ through functions of the form

$$\mu(\mathbf{x}) = \mu_0 + \mu_1 x_1 + \mu_2 x_2 + \ldots, \quad h(\sigma(\mathbf{x})) = \sigma_0 + \sigma_1 x_1 + \sigma_2 x_2 + \ldots, \quad \xi = \xi_0, \tag{13}$$

where $h(\cdot)$ is a link function in analogy to generalized linear models (Fahrmeir and Tutz, 2001).

The shape parameter is very sensitive to sampling uncertainty. Including covariates to the shape parameter largely increases the uncertainty not only of the shape parameters estimate itself, but also of the other parameters (Friederichs et al., 2009). The benefit in turn is generally small. For this reason, the shape parameter is kept independent of covariates. However, in a forecasting procedure, alike the other parameters, the shape parameter is estimated on a training data set and used to provide out-of-sample prediction. This out-of-sample prediction and verification is realized by a cross-validation procedure (see section 3.4).

As an alternative forecast, we apply a standard approach in meteorology and wind engineering, a generalized linear model (GLM, e.g., McCullagh and Nelder, 1999) for the gust factor. The gust factor G is defined as

$$G = \frac{y_{fx}}{x_{ff}} - 1, \tag{14}$$

where x_{ff} denotes the mean wind speed observation and y_{fx} the peak wind speed observation. The gust factor is generally assumed to follow a lognormal distribution, while the most simple approaches assume a constant gust factor. We base our approach on the work of Weggel (1999) and Jungo et al. (2002) and use a GLM for lognormal residuals. This corresponds to a standard linear regression model for $\log G$ with the expected value being modeled as

$$\mathbb{E}[\log G] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots. \tag{15}$$

The GLM is estimated using iteratively reweighted least squares as provided by the `glm` function in `R` (R Development Core Team, 2010).

3.1 Data

The Royal Netherlands Meteorological Institute freely provides daily weather data from numerous observation locations in the Netherlands[1]. The meteorological station data contain observations of temperature, sunshine, cloud cover and visibility, air pressure, precipitation, mean wind, and peak wind speed for the specified station. We have chosen the station number 210, Valkenburg, with hourly observations from January 1, 2001 to July 1, 2009 which gives us a total of 3104 daily observations. The daily observed peak wind y_{fx} is measured as a 24-hour (00UTC-00UTC) maximum over the observed hourly wind gusts. The gust observations are measured in 1 m/s which makes the data quasi discrete. In order to obtain a more continuous spectrum of values, we added a uniformly distributed noise to the gust observations. The hourly mean wind speed is given by the mean wind speed during the 10-minute period preceding the time of observation. We process these data to daily mean wind x_{ff} and daily wind variance x_{ffVar} by taking the average and the variance, respectively, over the 24 hourly observations as above. A similar procedure yields the mean rain rate x_{rr} and the maximum rain rate x_{rMax} over 24 hours. Finally, we also consider the mean 24-hour pressure x_P and the 24-hour pressure tendency x_{dP} as potential covariates. The pressure tendency x_{dP} is estimated by fitting a linear trend function (i.e., linear function in time) to the 24 hourly pressure observations. The slope estimate yields an estimate of x_{dP}. Note that the covariates are normalized before entering as a covariate. Normalization is not necessary but makes the inference procedures more stable.

[1] `http://www.knmi.nl/climatology/daily_data/download.html`

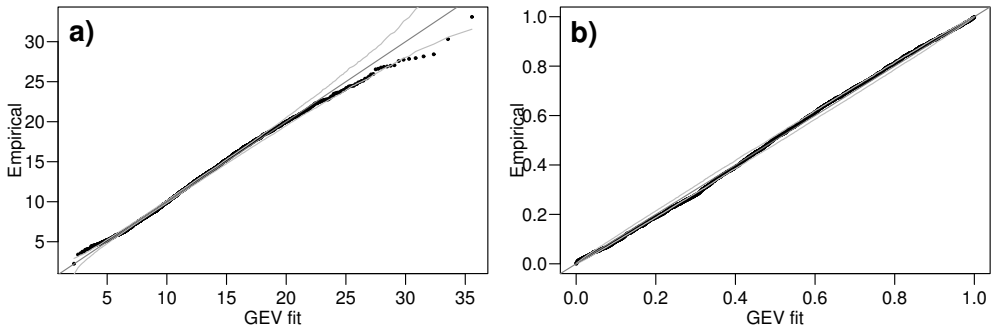

Figure 2: The goodness of fit of a seasonal GEV model applied to 24h peak wind speed observations at Valkenburg, the Netherlands from January 1, 2001 to July 1, 2009 measured by a QQ-plot (a) and a PP-plot (b). The dark gray line represents the seasonal GEV model estimates, the light gray lines indicate the 95% confidence interval derived by parametric bootstrapping, and the black dots denote the empirical estimates.

Peak wind speed refers to the maximum wind speed in a given time period. In operational terms, this refers to the highest 3-second average value recorded in a particular time period (e.g. 1 hour). Consequently, a block maxima approach using a GEV represents a natural approach to model the peak wind speed observations. Figure 2 represents a goodness of fit for a GEV model fitted to the daily peak wind speed observations in our data set using maximum likelihood estimation. In order to account for seasonal variations, we included the annual cycle in terms of a annual sine and cosine function as covariate to the location and scale parameter. Both parameters exhibit significant annual changes. The relation of the shape parameter to the annual cycle is not significantly different from zero, neither any relation to a half-annual sine or cosine function. Although significant, the annual cycle accounts for only about 5% of the variance in the daily peak wind speed observations. The seasonally varying GEV exhibits a shape parameter of about $\hat{\xi} = -0.022$ with a standard error of 0.014 and provides an overall acceptable fit even though it seems to slightly overestimate high quantiles, see Figure 2(a).

3.2 Optimum score estimation

In our frequentist approach, we apply both classic maximum likelihood estimation (GEV-MLE) and minimum CRPS estimation (GEV-CRPS) for the parameter estimation of the predictive distribution function. For the minimum CRPS estimation, we minimize the function

$$\sum_i S_{CRP}(F_i, y_{\mathbf{f}\mathbf{x},i})$$

where the index runs over the training set, S_{CRP} is given by either (9) or (10) depending on the value of the shape parameter ξ, and F_i is the predictive distribution for a covariate vector $\mathbf{x}_i$. The predictive distribution for Y is a GEV with

$$P(Y \leq y|\mathbf{x}) = F_{GEV}(y; \mu(\mathbf{x}), \sigma(\mathbf{x}), \xi), \tag{16}$$

where the parameters $\mu(\mathbf{x}), \sigma(\mathbf{x})$, and ξ are given by (13). Here, we apply both the identity and the logarithmic link function for the scale parameter σ. The latter ensures a positive scale parameter

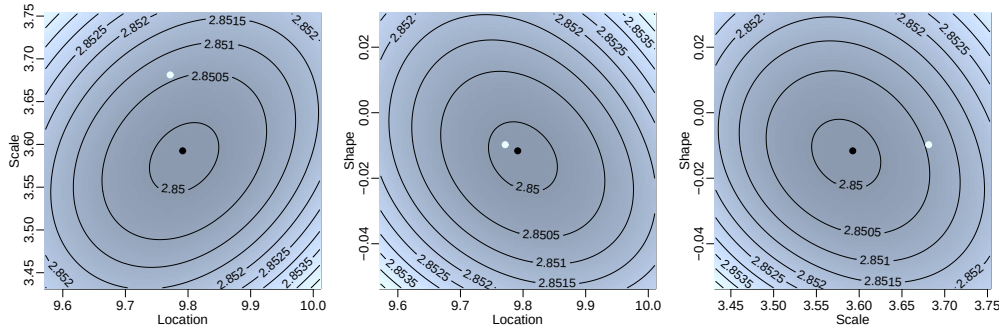

Figure 3: The in-sample ignorance score for the stationary GEV over the full data set. The plots show changes in the score when the location μ_0, scale σ_0, and shape ξ_0 are varied pairwise while the remaining parameter is fixed at the MLE. The black dots indicate the MLE and the white dots the minimum CRPS estimate.

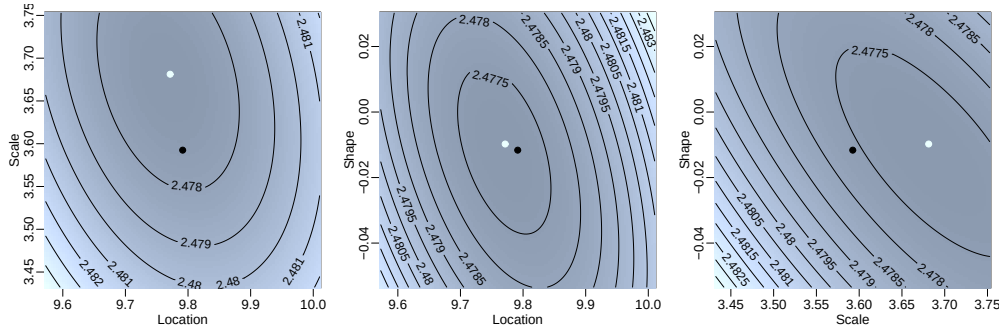

Figure 4: The in-sample CRPS for the stationary GEV over the full data set. The plots show changes in the score when the location μ_0, scale σ_0, and shape ξ_0 are varied pairwise while the remaining parameter is fixed at the minimum CRPS estimate. The white dots indicate the minimum CRPS estimates and the black dots the MLE.

and guarantees valid predictions.

Numerical optimization was first performed using the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno (BFGS) and the Nelder-Mead methods implemented in `R` (R Development Core Team, 2010) in parallel and the best estimates in terms of optimum score were used. However, as we included more covariates, the optimization became less stable. For that reason, all optimum score estimates are derived using a Simulated Annealing algorithm (SANN). The SANN method in `R` by default uses a variant of Simulated Annealing given in Belisle (1992) and is useful for getting good estimates on rough surfaces. Although the SANN algorithm is slower than the BFGS and the Nelder-Mead, it provides more robust estimates.

In order to compare the maximum likelihood and the minimum CRPS estimation, we investigate two-dimensional surfaces of the in-sample ignorance score and the CRPS over the entire data set. The link function for the scale parameter σ is the identity. Figure 3 shows two-dimensional surfaces of the log-likelihood function, or the ignorance score, for pairwise combinations of μ_0, σ_0, and ξ_0. In each plot, the other parameter is fixed at its respective maximum likelihood estimate (MLE). The log-likelihood function is displayed over a range that covers the parameter estimate

plus-minus three standard errors as derived from the profile likelihood (Coles, 2001). Similar two-dimensional surfaces are displayed for the CRPS in Figure 4, where in each plot, the other parameter is fixed at its respective minimum CRPS estimates. The CRPS is displayed over the same parameter range as the log-likelihood function in Figure 3. The MLE and the minimum CRPS estimates differ only slightly, or by less than two standard errors for all parameters (cf. the black and white dots in Figure 3 and 4). The pairwise dependence of changes in the parameter values on the function surfaces is quite different for the two functions. It is particularly strong for the CRPS where the dependence is negatively oriented. Further, the CRPS function is less sharp than the log-likelihood function in the direction of ξ_0.

3.3 Bayesian forecasting framework

Let $l(\mathbf{y}_{fX}|\boldsymbol{\theta}, \mathbf{X})$ denote the likelihood function under the GEV model in (7), where $\mathbf{y}_{fX}$ are the observed peak winds over the training set, $\mathbf{X}$ is the matrix of corresponding covariates, and $\boldsymbol{\theta}$ is a vector of the model parameters in (13) with a logarithmic link function for the scale parameter. Given a prior distribution $f(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, the joint posterior distribution of the parameters given the data is

$$f(\boldsymbol{\theta}|\mathbf{y}_{fX}, \mathbf{X}) \propto l(\mathbf{y}_{fX}|\boldsymbol{\theta}, \mathbf{X}) f(\boldsymbol{\theta}).$$

The predictive density for a new observation y' with a covariate vector $\mathbf{x}'$ is given by

$$f(y'|\mathbf{x}', \mathbf{y}_{fX}, \mathbf{X}) = \int l(y'|\boldsymbol{\theta}, \mathbf{x}') f(\boldsymbol{\theta}|\mathbf{y}_{fX}, \mathbf{X}) \mathrm{d}\boldsymbol{\theta}. \tag{17}$$

While this density is generally not a GEV density, it has the advantage that it reflects uncertainty both in the model and in the future observations, see e.g. Coles (2001). For the GEV model, the integral in (17) is usually intractable. However, the density may easily be approximated by a large sample from the predictive distribution, see Hoff (2009, section 4.3).

We use non-informative independent normal priors for the parameters and set $\mu_j, \sigma_j \sim N(0, 10^4)$ for all j, and $\xi_0 \sim N(0, 10^2)$. For stationary GEV distributions, it is common to use the logarithmic link function for the scale parameter, see e.g. Stephenson and Tawn (2004) and Galiatsatou et al. (2008). We follow their practice and only use the logarithmic link function for this part of the analysis. Our priors correspond to setting $\mu_0, \log(\sigma_0) \sim N(0, 10^4)$ and $\xi_0 \sim N(0, 10^2)$ in the stationary case. Stephenson and Tawn (2004) and Galiatsatou et al. (2008) argue that, in the stationary case, a trivariate normal distribution would be preferred here, as a negative dependence between σ_0 and ξ_0 is a priori expected. As our inference is based on a relatively large data set, we refrain from this given the computational complexity that such a prior would induce in the non-stationary case. We then apply a Metropolis within Gibbs algorithm to obtain samples from the marginal posterior distributions using normal distributions with a small variance as proposal distributions for the updates. For each parameter estimation, we run 100.000 iterations of the Metropolis within Gibbs algorithm with a burn-in period of 25.000 iterations. Classical diagnostics such as traceplots and running mean plots (not shown) show good mixing and indicate that this is sufficient for convergence.

In this case study, we consider six possible covariates and each of these can influence both the location and the scale parameter. We thus have a total of 2^{12} possible models in our model

Table 1: Average posterior inclusion probabilities for the covariates in a Bayesian model selection framework over all years. The probabilities are given in percentages.

Variable	Location	Scale
Mean wind speed (*ff*)	100.0	100.0
Wind variance (*ffVar*)	100.0	100.0
Mean rain rate (*rr*)	0.1	0.0
Maximum rain rate (*rMax*)	100.0	100.0
Pressure (*P*)	100.0	10.8
Pressure tendency (*dP*)	100.0	0.0

space. Besides considering specific models, we also perform a variable selection procedure over the entire model space. That is, we write each regression coefficient μ_j for $j \geq 1$ as $\mu_j = z_j \nu_j$, where $z_j \in \{0, 1\}$ and $\nu_j \in \mathbb{R}$ such that the z_j's indicate which regression coefficients are non-zero. The regression equation for μ becomes

$$\mu(\mathbf{x}) = \mu_0 + z_1\nu_1 x_{\textit{ff}} + z_2\nu_2 x_{\textit{ffVar}} + z_3\nu_3 x_{\textit{rr}} + z_4\nu_4 x_{\textit{rMax}} + z_5\nu_5 x_{\textit{P}} + z_6\nu_6 x_{\textit{dP}},$$

and similarly for $\log(\sigma(\mathbf{x}))$. Each model indication parameter z_j is a priori equal to either 0 or 1 with probability $1/2$, while the regression parameters ν_j have independent $N(0, 10^4)$ priors as before. Again, the parameters are updated iteratively using a Metropolis within Gibbs updating scheme.

Our algorithm is equivalent to a reversible jump MCMC algorithm (Green, 1995) and the posterior inclusion probability of a covariate x equals the average value of the corresponding indicator variable z over the posterior sample. A related Bayesian inference framework is proposed in El Adlouni and Ouarda (2009) where the authors suggest a birth-death MCMC procedure for covariate selection in generalized extreme value models and apply their framework to annual maximum precipitation data. However, the birth-death MCMC algorithm is quite complex to implement compared to the fairly simple regression variable selection method described in Hoff (2009) which we have applied. For more details on the regression variable selection algorithm and Bayesian inference in general, see Hoff (2009).

3.4 Model selection

In order to assess the predictive performance of our prediction approaches, we pursue a common approach in atmospheric science and use an out-of-sample verification. The prediction is performed in a cross-validation manner, in that we iteratively leave out one year of data, estimate the parameters of the statistical models based on the remaining data, and then perform an out-of-sample prediction. In this way, independent predictions for the complete time series are generated and used for verification and model selection.

Average posterior inclusion probabilities for the covariates under the Bayesian regression variable selection framework are given in Table 1. Three covariates, mean wind speed (*ff*), wind variance (*ffVar*), and maximum rain rate (*rMax*) have very high inclusion probabilities for both the

Table 2: Average predictive performance of non-stationary peak wind speed predictions at Valkenburg from January 1, 2001 to July 1, 2009. The performance is measured in terms of the continuous ranked probability score (CRPS), which is given in meters per second, and the ignorance score (IGN). The covariates considered here are mean wind speed *ff*, wind variance *ffVar*, mean rain rate *rr*, maximum rain rate *rMax*, pressure *P*, and pressure tendency *dP*. VS stands for Bayesian variable selection approach. The link function $h(\sigma)$ used for the scale parameter is indicated in parenthesis and the optimal score in each column is indicated in bold.

Covariate(s)	None	*ff*	*ff* *ffVar*	*ff* *ffVar* *rMax*	*ff* *ffVar* *rMax* $P^{\mu \text{ only}}$ $dP^{\mu \text{ only}}$	VS
CRPS (m/s)						
GEV-MLE (id)	**2.48**	**1.07**	0.86	0.82	0.80	—
GEV-MLE (log)		**1.07**	0.86	0.82	0.80	—
GEV-CRPS (id)	**2.48**	**1.07**	0.86	**0.81**	**0.79**	—
GEV-CRPS (log)		**1.07**	0.86	0.82	0.80	—
GEV-Bayes (log)	2.49	1.08	**0.84**	**0.81**	**0.79**	**0.80**
lognormal GLM	1.37	1.09	0.95	0.92	0.90	—
IGN						
GEV-MLE (id)	**2.85**	**2.00**	**1.78**	**1.73**	**1.71**	—
GEV-MLE (log)		**2.00**	**1.78**	**1.73**	1.72	—
GEV-CRPS (id)	**2.85**	**2.00**	**1.78**	**1.73**	1.72	—
GEV-CRPS (log)		**2.00**	**1.78**	1.74	1.73	—
GEV-Bayes (log)	**2.85**	**2.00**	**1.78**	**1.73**	**1.71**	**1.71**
lognormal GLM	2.20	2.03	1.90	1.86	1.85	—

location parameter μ and the scale parameter σ in (13). In addition, pressure (*P*) and pressure tendency (*dP*) also have high posterior inclusion probabilities for the location parameter μ. Note that the inclusion probabilities are calculated after the burn-in period has been removed. Especially for the low inclusion probabilities, the chains show a high degree of mixing in the burn-in period. Although seasonality may be captured by the existing covariates, we have further investigated adding annual cycles to the model. The results indicate that a small improvement in predictive performance (i.e. of the order of 0.02) may be obtained by including an annual cycle in the location parameter. Consequently, only a small fraction of seasonality is left unexplained when no annual cycle is concidered explicitly. For clarity of exposition, we omit this factor in the following.

Table 2 shows the average CRPS and ignorance score over all days in the test set for all the prediction methods considered in this study and different sets of covariates for each method. To calculate the CRPS in Table 2, we have applied (9) and (10) for the GEV models while for the Bayesian method, we have applied the approximation in (3). Since the logarithm of the gust factor $log(G)$ follows a normal distribution, the CRPS of the lognormal GLM is calculated using the closed-form expression in Gneiting et al. (2005). To calculate the ignorance score for the Bayesian

method, we obtain a non-parametric density estimate for the predictive density based on a large sample from the posterior predictive distribution. The ignorance score for the lognormal GLM reads $S_{IGN}(f_{\mathcal{N}}, \log(G)) = -\log(f_{\mathcal{N}}(\log(G))/(y - \mathit{ff})dy)$, where y is the respective gust value and $f_{\mathcal{N}}$ is the predictive normal distribution of $\log(G)$. Note that the scores are calculated on cross-validated predictions (Sec. 3.4), in order to provide an out-of-sample prediction and verification. Further, we estimate the sampling uncertainty of the CRPS and ignorance score via the bootstrap method (Efron and Tibshirani, 1993). To that end, we recalculated the scores by resampling the prediction-observation pairs using the bootstrap method with replacement. The standard deviations of the CRPS and ignorance score within a 1000 member bootstrap sample vary between 0.01 and 0.03 for all the GEV methods.

Including covariates in the prediction improves the predictive performance significantly compared to using a stationary model. The decrease in CRPS amounts to about 1.4 (1.7) when including *ff* (*ff, ffVar, rMax, P, dP*). With respect to a sampling uncertainty of about 0.03 this decrease is highly significant. For the models compared here, all method show the highest predictive skill when only covariates with high posterior inclusion probability are included. Including *P* and *dP* in the scale parameter and *rr* in both the location and the scale parameter, or a subset of these, does not provide significant improvements. Although the lognormal distribution seems to provide a better fit to the data in the stationary case, the non-stationary GEV models clearly outperform the lognormal GLM except if only *ff* is used as covariate. The improvement of the CRPS amounts to about 0.1 which is well above the sampling uncertainty of the CRPS. The differences between the various GEV approaches are more subtle and not very significant. For the frequentist methods, the identity link for the scale parameter yields slightly higher skill than the logarithmic link, especially in terms of the ignorance score for the best non-stationary model. For the more complex models, with three covariates or more, the optimum CRPS methods performs minimally better than the maximum likelihood methods in terms of the CRPS and the other way around for the ignorance score. The best scores are generally obtained within the Bayesian framework.

3.5 Predictive performance

Based on the results in the previous section, we now focus on the non-stationary GEV with location parameter

$$\mu(\mathbf{x}) = \mu_0 + \mu_1 x_{\mathit{ff}} + \mu_2 x_{\mathit{ffVar}} + \mu_3 x_{\mathit{rMax}} + \mu_4 x_{\mathit{P}} + \mu_5 x_{\mathit{dP}} \tag{18}$$

and scale parameter

$$h(\sigma(\mathbf{x})) = \sigma_0 + \sigma_1 x_{\mathit{ff}} + \sigma_2 x_{\mathit{ffVar}} + \sigma_3 x_{\mathit{rMax}}, \tag{19}$$

where h is the identity link for the optimum score estimation and the logarithm for the Bayesian method. Although the differences are barely significant in table 2, there is evidence that the identity link model provides slightly better skill compared to the logarithm when using optimum score estimation. However, the following results also apply to the logarithm link model. For comparison, we also investigate the performance of the lognormal GLM method. Verification is then based on a variety of diagnoses.

The calibration, or the reliability, of the predictions is assessed by using residual quantile plots based on the standard Gumbel distribution, see Figure 5. The first column of Figure 5 shows the

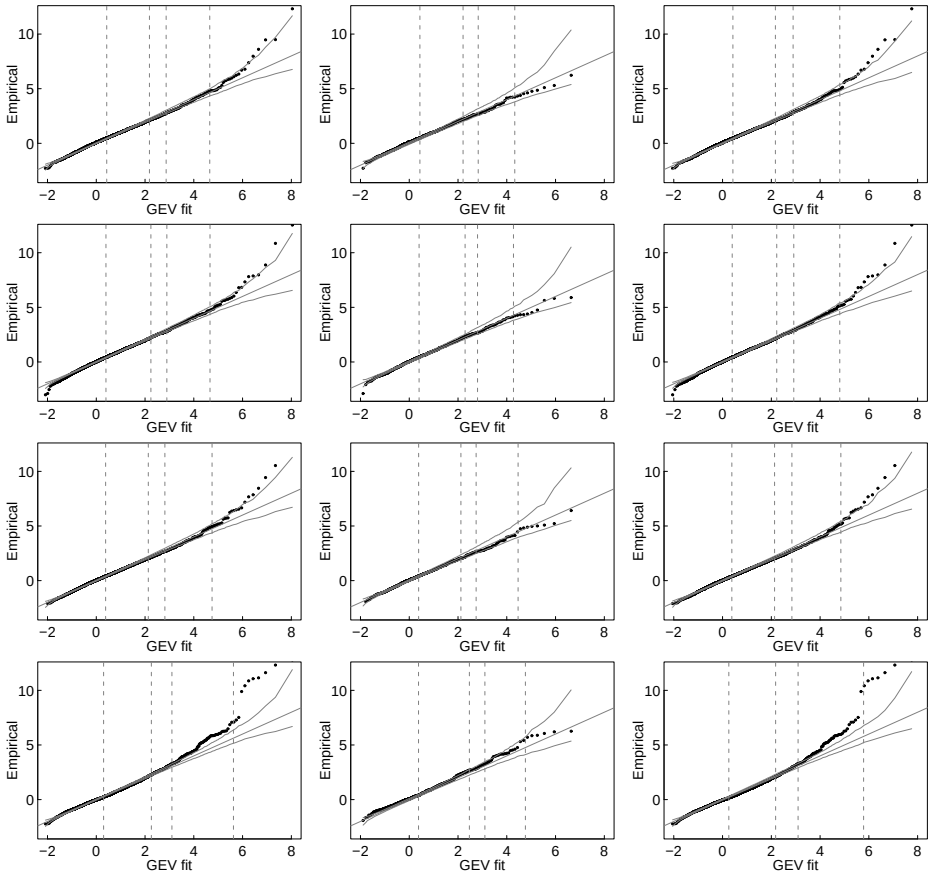

Figure 5: Residual quantile plots on Gumbel scale for models with the location parameter in (18) and the scale parameter in (19). The prediction methods compared here are the GEV-MLE method with identity link (1st row), the GEV-CRPS method with identity link (2nd row), the Bayes method (3rd row), and the lognormal GLM (4th row). The first column shows the QQ-plots over all forecasts, the second column shows forecasts with $x_{ff} > q_{.75}$, and the third column shows forecasts with $x_{ff} \leq q_{.75}$. The vertical lines indicate (from left to right) the residual 0.5, 0.9, 0.95, and 0.99 quantiles.

residual quantile plots (section 2.1) for the complete set of observations. For the lowest 99% of the residuals, the bulk of the residuals lies within the 95% uncertainty band of the residual distribution for all three GEV methods. However, the models show a significant underestimation of the highest 1% of the residuals. The underestimation is even more prominent for the logarithmic GLM where significant underestimation is observed for the highest 5% of the residuals. An additional stratification is based on the covariate x_{ff}, where predictions with below normal mean wind speed ($x_{ff} \leq q_{.75}$) and above normal mean wind speed ($x_{ff} > q_{.75}$) are considered separately, where $q_{.75}$ is the 0.75 quantile of the x_{ff} observations. In situations where the mean wind x_{ff} is above its 0.75 quantile, all models provide calibrated predictive distributions (Fig. 5, second column). The uncalibrated large residuals only occur during weak wind situations (Fig. 5, third column). Since strong gusts are very unlikely during these weather situations, this miscalibration is probably not highly relevant for the gust prediction.

The three GEV methods in Figure 5 thus all appear to be fairly well calibrated and we can

Table 3: Average Brier score at three thresholds and quantile score at four quantiles of peak wind speed predictions at Valkenburg from January 1, 2001 to July 1, 2009. The GEV methods apply the location parameter in (18) and the scale parameter in (19). The link function for the scale parameter is indicated in parenthesis. The best performance in each column is indicated in bold.

	Brier Score (1/100)			Quantile Score (1/10 m/s)			
	14m/s	18m/s	25m/s	0.75	0.9	0.95	0.99
GEV-MLE (id)	6.19	3.27	0.41	4.86	2.97	1.88	0.57
GEV-CRPS (id)	6.18	**3.21**	**0.39**	**4.81**	**2.95**	**1.87**	**0.56**
GEV-Bayes (log)	**6.16**	3.24	0.43	4.87	2.98	**1.87**	0.58
lognormal GLM	6.55	3.70	0.66	6.73	3.47	2.20	0.69

Table 4: Average skill scores (in percentages) of peak wind speed predictions at Valkenburg from January 1, 2001 to July 1, 2009. The GEV methods apply the location parameter in (18) and the scale parameter in (19). The link function for the scale parameter is indicated in parenthesis. The skill scores for the thresholds is based on the Brier score and the skill scores for the quantiles is based on the quantile score. The reference forecast is the stationary GEV-MLE method. The best performance in each column is indicated in bold.

	Threshold			Quantile			
	14m/s	18m/s	25m/s	0.75	0.9	0.95	0.99
GEV-MLE (id)	68.8	64.0	52.5	68.7	68.2	66.7	63.7
GEV-CRPS (id)	68.9	**64.7**	**54.6**	**69.0**	**68.4**	**67.0**	**64.6**
GEV-Bayes (log)	**69.0**	64.4	50.4	68.7	68.1	**67.0**	63.3
lognormal GLM	67.0	59.4	23.7	56.7	62.3	61.1	56.3

compare the sharpness of the predictive distributions in concurrence with the forecasting principle of Gneiting et al. (2007). The GEV-CRPS method yields the sharpest predictions with an average predictive standard deviation of $1.43\,(\pm 0.44)$, while this value is $1.48\,(\pm 0.57)$ for the Bayesian method and $1.51\,(\pm 0.48)$ for the GEV-MLE method. The values given in the parentheses are the standard deviations of the daily predictive standard deviations. The large day-to-day variation in the sharpness is to be expected as different levels of uncertainty are associated with different prediction scenarios. In general, the forecasts are more uncertain for higher expected peak wind speeds. An alternative approach to assess the sharpness is to calculate the average width of symmetric prediction intervals, see e.g. Gneiting et al. (2005) and Raftery et al. (2005). Similarly, the average coverage of the prediction intervals may be used to assess calibration.

Table 3 shows the Brier scores at three high thresholds and the quantile scores for four high quantiles for our four methods. Wind speeds of $14 - 18$m/s are generally considered as near gale, wind speeds of $18 - 25$m/s are defined as gale and strong gale, while storm values are 25m/s and above. The corresponding skill scores are shown in Table 4 where the reference forecast is the stationary GEV-MLE method. The standard error for the Brier score is $(0.31, 0.25, 0.09)$ for the thresholds $(14, 18, 25)$ and the standard error for the quantile score is about 0.10 for each quantile. These scores measure predictive performance at certain points of the predictive distribution or the

quantile function, and may indicate local deficiencies. The lognormal GLM provides significantly less skill for all thresholds and probabilities, except for the 0.99 quantile where the difference is no longer significant. All methods based on a non-stationary GEV distribution perform similarly, with the GEV-CRPS methods showing marginally better performance than the other two methods.

Note that the scores in Table 3 cannot be compared across columns as these are based on different subsets of the data. The apparent improvement in the scores for higher thresholds/quantiles is merely an effect of the decreasing data set on which the values are based. This is evident if the results of Table 3 are compared to the results on Table 4. As the same reference forecast is used for all the columns in Table 4, we can compare the results across columns relative to the reference forecast. Here, we see that the performance of all the methods decreases somewhat with higher thresholds/quantiles compared to the stationary GEV-MLE reference method.

Alternatively, weighted proper scoring rules may be applied to focus on areas of special interest in the predictive distribution. Diks et al. (2011) consider weighted versions of the ignorance score while Gneiting and Ranjan (2011) discuss approaches to weight the CRPS. The weight function can here either be based on quantiles using the representation in (6) or it can be based on thresholds using the representation in (1). Note that the weight functions must be defined independent of the observed value to uphold propriety (Gneiting and Ranjan, 2011).

3.6 Parameter estimation

Verification in terms of scores and residual quantile plots provides average performance of the prediction method. Another aspect is the variance and covariances of the parameter estimates. Bayesian prediction provides estimates of the posterior distribution of the parameters, whereas optimum score estimation only gives approximate covariances based on the profile score function (i.e. profile likelihood (Coles, 2001) in maximum likelihood estimation). An indication of uncertainty in the parameter estimation is provided by the variability of the estimates over the different training data set used in cross-validation. In the cross-validation procedure we successively remove one year of data, which provides us with nine estimates of each parameter for each method.

Figure 6 shows the parameter estimates for the yearly shape parameter ξ under the three GEV methods discussed in the previous section. The 90% posterior intervals obtained with the Bayesian method are fairly constant between years, while there is a large variation in the estimated values under both GEV-MLE and GEV-CRPS. Furthermore, the standard errors are significantly greater for the GEV-CRPS method. This is in accordance with Fig. 1 where the minimum of the CRPS is less sensitive to changes in ξ and the discussion related to Fig. 4. The MLE and CRPS optimum score estimates of ξ are uncorrelated which suggests that the variability in the estimates is not due to interannual changes in the data. The majority of the estimates are significantly negative, indicating that the non-stationary GEV models are of Weibull type, in contrast to the goodness of fit analysis performed in Section 3.1, where $\hat{\xi}$ is only slightly negative.

The parameter estimates for the location coefficients in (18) are shown in Figure 7. For clarity of the presentation and as the 90% posterior intervals are very similar for each run of the Bayesian method, we only show the 90% posterior interval of the joint posterior sample from all years. In general, there is no apparent correlation between the maximum likelihood estimates and the optimum CRPS estimates. Furthermore, the estimates for both methods are quite variable for different training sets and this variability cannot be explained solely by the uncertainty in the estimation which is of similar magnitude for both methods and across training sets. All three methods display

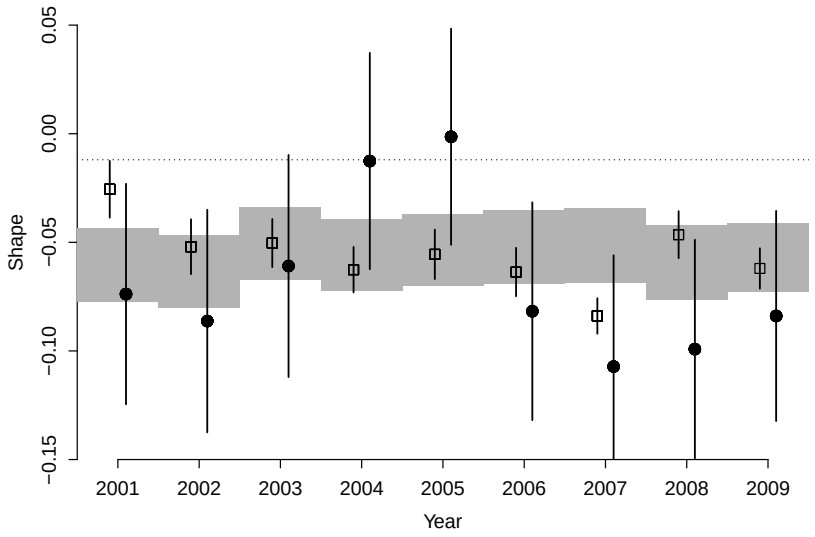

Figure 6: The estimated yearly shape parameters for the GEV methods using the location parameter in (18) and the scale parameter in (19). The gray boxes indicate the 90% posterior intervals for for ξ under the Bayesian model, the black dots ($\bullet$) indicate the parameter estimates under the GEV-CRPS(id) model, and the squares ($\square$) indicate the parameter estimates under the GEV-MLE(id) model. The vertical lines indicate the respective standard errors. For comparison, the shape parameter of the observations estimated via MLE, $\hat{\xi} = -0.013$, is indicated with a dotted line.

the largest uncertainty in the estimate for the intercept coefficient. All methods show a strong positive dependence of the GEV location parameter on *ff* and *ffVar*, a weaker dependence on *rMax*, and a negative dependence on pressure p and pressure tendency *dP*.

The Bayesian posterior distributions for the scale coefficients in (19) cannot be directly compared to the corresponding estimates under the frequentist approaches as we have applied the identity link for the frequentist approaches while we use a logarithmic link for the Bayesian method. As for the location coefficients, no apparent correlation can be found between the MLE coefficient estimates and the minimum CRPS estimates and the variability between training sets is generally greater than expected based on the estimation uncertainty alone. Like the location parameter, the non-stationary GEV scale parameter positively depends on *ff*, *ffVar*, and *rMax*. Hence, predictive uncertainty increases for large expected peak wind speed.

4 Discussion

As e.g. Dawid (1984) and Gneiting (2008) have argued before us, predictions for uncertain events ought to include an estimate of the associated uncertainty. This is easily obtained within a probabilistic framework where the predictions take the form of probability distributions. Proper scoring rules such as the Brier score or the CRPS are widely used to assess the predictive skill in probabilistic weather forecasting, see e.g. Wilks (2011). In this paper, we discuss how proper scoring rules may be used for out-of-sample model selection and verification for extreme value distributions. We present a closed-form expressions of the CRPS for this class of distributions. With these expressions at hand, the CRPS may be used for optimum score estimation as an alternative

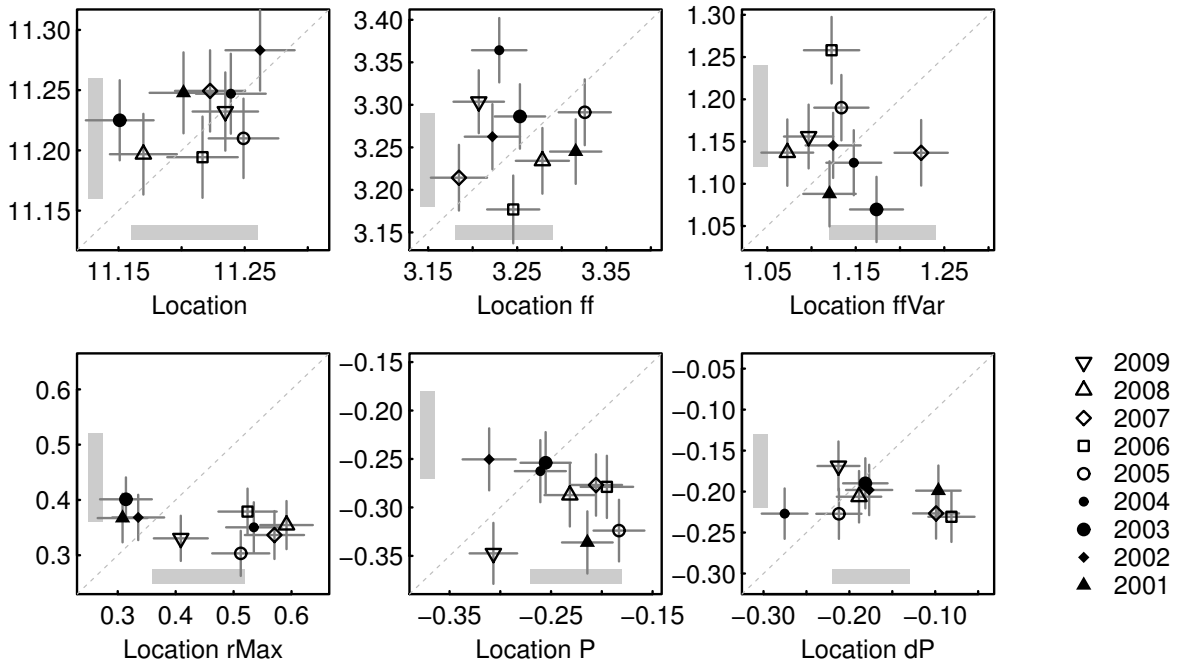

Figure 7: Optimum score estimates of yearly location parameters for the GEV methods using the location parameter in (18) and the scale parameter in (19). The dots represent the maximum likelihood estimates (x-axis) and the optimum CRPS estimates (y-axis) for the different years. The gray lines indicate the respective standard errors. The gray bars on each axis indicate the 90% posterior intervals for the parameters under the Bayesian model.

to maximum likelihood estimation or Bayesian inference.

In a case study, we compare various approaches to derive non-stationary distributions for daily peak wind speed observations. The non-stationarity is built in by including covariates, i.e. meteorological parameters that are observed together with the gust measurement. Our competing methodologies comprise a lognormal GLM for the wind gust factor and non-stationary GEV models for the daily wind gusts. The non-stationary GEV approaches provide significantly higher skill than the corresponding lognormal GLM. This confirms findings in Friederichs et al. (2009) that extreme value theory provides an appropriate and theoretically consistent statistical model for wind gusts.

Due to the close relationship between wind speed and peak wind, mean wind speed is the main covariate in our model. However, the predictive performance is significantly improved by also including information on additional weather variables. The most informative covariates besides mean wind speed are the variability of the mean wind speed, the maximum rain rate, and the pressure tendency throughout the day. With six potential covariates for both the location and the scale parameter, our model space includes a total of 2^{12} models. This large space can easily be searched using a Bayesian variable selection procedure which returns different covariate sets for the location and for the scale. Overall, the robustness and the predictive performance of the Bayesian framework is very good though it should be noted that the Bayesian inference is considerably slower than the frequentist methods.

Gneiting et al. (2005) compare minimum CRPS estimation and maximum likelihood estimation

in a prediction framework and show that the former returns forecasts with slightly better overall predictive performance and significantly improved calibration. Their results confirm well with our findings in that the optimum CRPS estimation outperforms the maximum likelihood approach in nearly all aspects of our verification even though the differences are sometimes small. A drawback of the minimum CRPS estimation is its weak discriminant power with respect to the GEV shape parameter.

In our case study, we have performed an extensive analysis of the properties of the various scoring rules and associated inference methods for the GEV distribution. The GEV is the self-evident distribution in the case of block maxima, i.e. in the case of peak wind speed. In the case of threshold excesses, a peak-over-threshold (POT) or Poisson point process approach would be more appropriate. We assume that similar results will hold for the POT approach using a GPD distribution and the closed-form expression given in this paper. Since the Poisson point process is generally represented in terms of GEV parameters, optimum score estimation using the CRPS should be possible as well.

Acknowledgments

Special thanks go to Michael Scheuerer for valuable discussions and comments on early versions of the manuscript. The authors also thank Tilmann Gneiting, Andreas Hense, Alex Lenkoski, and Michael Weniger for helpful discussions, and the associate editor and an anonymous reviewer for their useful comments. We acknowledge the support of the Volkswagen Foundation through the project “Mesoscale Weather Extremes - Theory, Spatial Modeling and Prediction (WEX-MOP)”.

References

Abramowitz M Stegun IA. 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York.

Anderson JL. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate* **9**: 1518–1530.

Beirlant J, Goegebeur Y, Segers J, Teugels J. 2004. *Statistics of Extremes*, Wiley, Chichester.

Belisle CJP. 1992. Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *J. Appl. Probab.* **29**: 885–895.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**: 1–3.

Bröcker J. 2009. Reliability, sufficiency, and the decomposition of proper scores. *Q. J. Roy. Meteor. Soc.* **135**: 1512–1519.

Bröcker J Smith LA. 2007. Scoring Probabilistic Forecasts: The Importance of Being Proper. *Weather Forecast.* **22**: 382–388.

Coles S. 2001. *An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics. Springer-Verlag, London.

Dawid AP. 1984. Statistical theory: The prequential approach (with discussion and rejoinder). *J. Roy. Stat. Soc. A* **147**: 278–292.

—. 2007. The geometry of proper scoring rules. *Ann. Inst. Statist. Math.* **59**: 77–93.

Diks C, Panchenko V, van Dijk D. 2011. Likelihood-based scoring rules for comparing density forecasts in tails. *J. Econometrics* **163**: 215–230.

Efron B Tibshirani RJ. 1993. *An Introduction to the Bootstrap*, Chapman & Hall.

El Adlouni S Ouarda TBMJ. 2009. Joint Bayesian model selection and parameter estimation of the generalized extreme value model with covariates using birth-death Markov chain Monte Carlo. *Water Resour. Res.* **45**: W06403.

Fahrmeir L Tutz G. 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer series in statistics, Springer, New-York.

Friederichs P, Goeber M, Bentzien S, Lenz A, Krampitz R. 2009. A probabilistic analysis of wind gusts using extreme value statistics. *Meteorol. Z.* **18**: 615–629.

Friederichs P Hense A. 2007. Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Weather Rev.* **135**: 2365–2378.

Galassi M, Davies J, Theiler J, Gough B, Jungman G, Alken P, Booth M, Rossi F. 2010. *GNU Scientific Library Reference Manual*, Network Theory Ltd, 1st ed.

Galiatsatou P, Prinos P, Sanchez-Arcilla A. 2008. Estimation of extremes: conventional versus Bayesian techniques. *J. Hydraul. Res.* **46**(S2): 211–223.

Gneiting T. 2008. Editorial: Probabilistic forecasting. *J. Roy. Stat. Soc. A* **171**: 319–321.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. B* **69**: 243–268.

Gneiting T, Larson K, Westrick K, Genton MG, Aldrich E. 2006. Calibrated probabilistic forecasting at the Stateline wind energy center: The regime-switching space-time method. *J. Am. Stat. Assoc.* **101**: 968–979.

Gneiting T Raftery AE. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **102**: 359–378.

Gneiting T, Raftery AE, Westveld AH, Goldman T. 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**: 1098–1118.

Gneiting T Ranjan R. 2011. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econ. Stat.* **29**(3): 411–422.

Good IJ. 1952. Rational Decisions. *J. Roy. Stat. Soc. B* **14**: 107–114.

Green P. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**(4): 711.

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA. 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q. J. Roy. Meteor. Soc.* **132**: 2925–2942.

Guttorp P Fuentes M. 2010. Extreme events in climate and weather – an interdisciplinary workshop. Available online at `http://www.birs.ca/workshops/2010/10w5016/report10w5016.pdf`.

Hamill TM Colucci SJ. 1997. Verification of Eta-RSM short-range ensemble forecasts. *Mon. Weather Rev.* **125**: 1312–1327.

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**: 559–570.

Hoff PD. 2009. *A First Course in Bayesian Statistical Methods*, Springer.

Jungo P, Goyette S, Beniston M. 2002. Daily wind gust speed probabilities over Switzerland according to three types of synoptic circulation. *Int. J. Climatol.* **22**: 485–499.

Laio F Tamea S. 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sc.* **11**: 1267–1277.

McCullagh P Nelder J. 1999. *Generalized Linear Models*, vol. 37 of *Monographs on Statistics and Applied Probability*, Chapman&Hall/CRC.

Murphy AH. 1973. Hedging and skill scores for probability forecasts. *J. Appl. Meteorol.* **12**: 215–223.

—. 1974. A sample skill score for probability forecasts. *Mon. Weather Rev.* **102**: 48–55.

—. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecast.* **8**: 281–293.

Murphy AH Winkler RL. 1987. A general framework for forecast verification. *Mon. Weather Rev.* **115**: 1330–1338.

R Development Core Team 2010. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Mon. Weather Rev.* **133**: 1155–1174.

Roulston M Smith L. 2002. Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130**: 1653–166.

Selten R. 1998. Axiomatic characterization of the quadratic scoring rule. *Exp. Econ.* **1**: 43–62.

Stephenson A Tawn J. 2004. Bayesian inference for extremes: accounting for the three extremal types. *Extremes* **7**: 291–307.

Thorarinsdottir TL Gneiting T. 2010. Predicting inflation: professional experts versus no-change forecasts. ArXiv:1010.2318v1 [stat.AP].

Unger DA. 1985. A method to estimate the continuous ranked probability score. *Preprints, Ninth Conf. on Probability and Statistics in Atmospheric Sciences, Virginia Beach, VA, Amer. Meteor. Soc.* : 206–213.

Weggel JR. 1999. Maximum daily wind gusts related to mean daily wind speed. *J. Struct. Eng. - ASCE* **125**: 465–468.

Wilks D. 2011. *Statistical Methods in the Atmospheric Sciences*, vol. 100 of *International Geophysics Series*, Academic Press, 3rd ed.

Appendix A: Derivation of the CRPS for a GEV

For the deviation of the CRPS under the GEV, we apply the representation in (6), where the CRPS is presented as an integral over the quantile score in (5). We can write the quantile score as

$$S_Q^\tau(F,y) = \tau\big[y - F^{-1}(\tau)\big] - \mathbb{1}\{\tau \geq F(y)\}\big[y - F^{-1}(\tau)\big]$$

from which it is easily seen that

$$S_{CRP}(F,y) = y\big[2F(y)-1\big] - 2\int_0^1 \tau F^{-1}(\tau)d\tau + 2\int_{F(y)}^1 F^{-1}(\tau)d\tau. \tag{20}$$

The quantile function of the GEV is given by

$$F_{GEV}^{-1}(\tau) = \begin{cases} \mu - \frac{\sigma}{\xi}\big[1-(-\log\tau)^{-\xi}\big], & \xi \neq 0 \\ \mu - \sigma\log\left(-\log\tau\right), & \xi = 0 \end{cases}. \tag{21}$$

For a non-zero shape parameter $\xi \neq 0$ and with $F = F_{GEV_{\xi\neq 0}}$, we obtain

$$2\int_0^1 \tau F^{-1}(\tau)d\tau = \mu - \frac{\sigma}{\xi} + 2\int_0^1 \tau\big(-\log\tau\big)^{-\xi}d\tau \tag{22}$$

and

$$2\int_{F(y)}^1 F^{-1}(\tau)d\tau = 2\Big[\mu - \frac{\sigma}{\xi}\Big]\big[1-F(y)\big] + 2\frac{\sigma}{\xi}\int_{F(y)}^1 \big(-\log\tau\big)^{-\xi}d\tau. \tag{23}$$

To solve the two remaining integrals in (22) and (23), we use that

$$\int \tau(-\log\tau)^{-\xi}d\tau = 2^{\xi-1}\Gamma_u(1-\xi, -2\log\tau)$$

and

$$\int (-\log\tau)^{-\xi}d\tau = \Gamma_u(1-\xi, -\log\tau),$$

where $\Gamma_u(a,\tau) = \int_\tau^\infty t^{a-1}e^{-t}dt$ is the upper incomplete gamma function. By combining (20), (22), (23), and the integral equations above, we get

$$S_{CRP}(F_{GEV_{\xi\neq0}},y) = \Big[y-\mu+\frac{\sigma}{\xi}\Big]\big[2F_{GEV_{\xi\neq0}}(y)-1\big] - \frac{\sigma}{\xi}\Big[2^\xi\Gamma(1-\xi) - 2\Gamma_l\big(1-\xi, -\log F_{GEV_{\xi\neq0}}(y)\big)\Big],$$

where $\Gamma_l(a,\tau) = \Gamma(a) - \Gamma_u(a,\tau)$ is the lower incomplete gamma function.

For a Gumbel type GEV with $\xi = 0$, similar calculations show that

$$\begin{aligned} S_{CRP}(F_{GEV_{\xi=0}},y) =& [y-\mu]\big[2F_{GEV_{\xi=0}}(y)-1\big] + 2\sigma\int_0^1 \tau\log(-\log\tau)d\tau \\ &- 2\sigma\int_{F_{GEV_{\xi=0}}}^1 \log(-\log\tau)d\tau. \end{aligned} \tag{24}$$

The first integral is solved as

$$\int \tau \log(-\log \tau) d\tau = \frac{1}{2}\Big[\tau^2 \log(-\log \tau) - Ei(2 \log \tau)\Big]$$

where $Ei(x) = \int_{-\infty}^{x} \frac{e^t}{t} dt$ is the exponential integral also given by

$$Ei(x) = C + \log(|x|) + \sum_{k=1}^{\infty} \frac{x^k}{k!\, k}$$

with $C \approx 0.5772$ being the Euler-Mascheroni constant. Further, we use

$$\int \log(-\log \tau) d\tau = \tau \log(-\log \tau) - Ei(\log \tau),$$

to solve the second integral in (24). This leads to a closed-form expression for the CRPS under a Gumbel-type GEV with

$$\begin{aligned}
&S_{CRP}(F_{GEV_{\xi=0}}, y) \\
&\quad = [y - \mu]\big[2F_{GEV_{\xi=0}}(y) - 1\big] \\
&\qquad + \sigma \lim_{\nu \to 1} \Big[\nu^2 \log(-\log \nu) - Ei(2 \log \nu)\Big] - \lim_{\eta \to 0} \Big[\eta^2 \log(-\log \eta) - Ei(2 \log \eta)\Big] \\
&\qquad - 2\sigma \lim_{\nu \to 1} \Big[\nu \log(-\log \nu) - Ei(\log \nu)\Big] - 2\sigma \Big[F_{GEV_{\xi=0}}(y) \frac{y - \mu}{\sigma} + Ei\big(\log F_{GEV_{\xi=0}}(y)\big)\Big] \\
&\quad = -(y - \mu) - 2\sigma Ei\big(\log F_{GEV_{\xi=0}}(y)\big) \\
&\qquad + \sigma \lim_{\nu \to 1} \Big[(\nu - 1)^2 \log(-\log \nu) + C - \log 2 + \sum_{k=1}^{\infty} \frac{(2 \log \nu)^k + 2(\log \nu)^k}{k! k}\Big] \\
&\quad = -\,(y - \mu) - 2\sigma Ei\big(\log F_{GEV_{\xi=0}}(y)\big) + \sigma[C - \log 2],
\end{aligned}$$

where we use that $\lim_{\eta \to -\infty} Ei(\eta) = 0$. The series expansion to compute $Ei(\log F_{GEV_{\xi=0}}(y))$ converges rapidly for $F_{GEV_{\xi=0}}(y)$ much greater than 0. However, it may fail for very small values of $F_{GEV_{\xi=0}}(y)$. We use the GNU Scientific Library (Galassi et al., 2010) to compute $Ei(\cdot)$.

Appendix B: Derivation of the CRPS for a GPD

The derivation of the closed-form expression of the CRPS for a GPD is very similar to the derivations for a GEV in Appendix A with significantly simpler calculations. The GPD quantile function is given by

$$F^{-1}_{GPD}(p) = \begin{cases} u - \frac{\sigma_u}{\xi} + \frac{\sigma_u}{\xi}(1 - p)^{-\xi}, & \xi \neq 0 \\ u - \sigma_u \log(1 - p), & \xi = 0 \end{cases}.$$

By applying the representation of the CRPS given in (20), we get

$$\begin{aligned}
S_{CRP}(F_{GPD_{\xi\neq 0}}, y) &= \Big[y - u + \frac{\sigma_u}{\xi}\Big]\big[2F_{GPD_{\xi\neq 0}}(y) - 1\big] \\
&\quad - \frac{2\sigma_u}{\xi}\Big[\int_{F_{GPD_{\xi\neq 0}}(y)}^{1}(1-\tau)^{-\xi}d\tau - \int_0^1 \tau(1-\tau)^{-\xi}d\tau\Big] \\
&= \Big[y - u + \frac{\sigma_u}{\xi}\Big]\big[2F_{GPD_{\xi\neq 0}}(y) - 1\big] \\
&\quad - 2\frac{\sigma_u}{\xi(\xi-1)}\Big[\frac{1}{\xi-2} + \Big[1 - F_{GPD_{\xi\neq 0}}(y)\Big]\Big[1 + \xi\frac{y-u}{\sigma_u}\Big]\Big].
\end{aligned}$$

Here, the second integral can be solved using integration by parts. For $\xi = 0$, we obtain the following expression

$$\begin{aligned}
S_{CRP}(F_{GPD_{\xi=0}}, y) &= (y-u)\big[2F_{GPD_{\xi=0}}(y) - 1\big] \\
&\quad + 2\sigma_u\Big[\int_0^1 \tau\log(1-\tau)d\tau - \int_{F_{GPD_{\xi=0}}(y)}^{1}\log(1-\tau)d\tau\Big] \\
&= y - u - \sigma_u\Big[2F_{GPD_{\xi=0}}(y) - \frac{1}{2}\Big],
\end{aligned}$$

where we use the integral equation

$$\int \log(1-x)dx = x\big[\log(1-x) - 1\big] - \log(1-x)$$

and integration by parts.